SYSTEMS DESIGN DOCUMENT

GROUP MEMBERS.

1.Tumusiime Arnold

2.Muyanja Rodney

3.Tusiime Allan

4.Shyaka Stephen

* ***DATA PIPELINE DIAGRAM.***

Model Deployment

Data Wrangling

Data Modelling

Data Loading

* ***DESCRIPTION OF KEY COMPONENTS IN THE DATA PIPELINE DIAGRAM.***
* Data Loading.

This involves loading of data into the pipeline in a particular format from the data source. Our dataset is collected from kaggle.com and is in CSV format.

This involves the use of a python library known as pandas which is imported as pd for usability purposes in the code. This library is used to load the dataset into the pipeline by running a method called pd.read\_csv(url). This particular library is preferred because of its immense ability to load various datasets in various formats.

* Data Wrangling.

This involves cleaning of data before use. For our particular dataset we carried out label encoding so as to assign digits to strings, feature scaling, and also dropped columns with missing values and other irrelevant columns.

This also involves the use of the pandas python library. Here various methods are used such as the pd.get\_dummies for label encoding, pd.concat for combining dataframes, pd.drop for dropping columns. This particular library is preferred because of it contains all these various methods to clean datasets in ones preferred way.

* Data Visualization.

This involves visualizing various features within the dataset so as to perform analysis in the dataset features. This also helps in performing feature correlation so as to know what features to be dropped.

This component involves the use of the matplotlib and seaborn python libraries. These libraries are used to create various visual formats that are based upon to create the analysis on the dataset.

* Data Modelling.

This involves creating models and training them to correlate the data with the outcomes. Various models are used but the one with the best accuracy score is chosen to be deployed. These Models include

* LogisticRegression model.
* SVC model.
* KNeighborsClassifier model.
* DecisionTreeClassifier model.

This component involves the use of the sklearn, numpy, and pandas python libraries. The numpy library is preferred because it is able to manipulate various mathematical functions, and the sklearn library is preferred because it is able to create models using their respective methods.

* Model Deployment.

This involves deployment of the models with the highest score after using the train data. The test data is the used by the selected models. This component involves the use of the sklearn, numpy, and pandas python libraries. The numpy library is preferred because it is able to manipulate various mathematical functions, and the sklearn library is preferred because it is able to create models using their respective methods.

* ***Visualization Components.***